TD Régression Logistique

On a relevé sur un échantillon de $n = 900$ individus, les personnes souffrant de problèmes de circulation sanguine et la consommation journalière de tabac (en nombre de paquets). On a obtenu le tableau suivant :

Etat de santé \conso.	0	1	≥ 2
Malade	40	70	200
Sain	260	230	100

1. Faites un test pour tester la dépendance entre la consommation de tabac et la présence de problèmes de circulation.

On veut appliquer le modèle de régression logistique. On note Y la variable qui vaut 1 si le sujet est malade, 0 s'il est sain et les probabilités conditionnelles

$$\pi_i = \mathbb{P}(Y = 1|C_i), \;\; i = 0, 1, 2\,,$$

où C_i désigne la population des individus fumant i paquets par jour.

2. Choisissez un profil de référence et donnez le codage d'un individu
   — malade et non fumeur ;
   — sain et fumant plus de 2 paquets par jour.
3. Donnez pour tout $i \in \{0, 1, 2\}$, l'expression de π_i dans le cadre d'un modèle de régression logistique.
4. Exprimez les paramètres $(\beta_i, i \in \{0, 1, 2\})$ en fonction des probabilités conditionnelles $(\pi_i, i \in \{0, 1, 2\})$.
5. Calculez les probabilités empiriques d'être malade sachant chaque classe.
6. Calculez la cote d' "être malade" pour chaque classe, ainsi que le rapport des cotes.
7. Exprimez, puis calculez les estimateurs $\hat{\beta}_i$.
8. Expliquez les effets des deux paramètres sur la variation des probabilités d'être malade.
9. Calculez la déviance du modèle total et du modèle vide. Faites un test de la pertinence de la régression.
10. Quels sont les coefficients significativement différents de 0 ? Donnez un intervalle de confiance de coefficients de sécurité 95% pour π_1.
11. Quelle est la table de confusion du modèle ?
12. Tracez la courbe ROC du modèle.